DOCUMENT RESUME

ED 169 096 TM 008 492

AUTHOR	Epstein, Kenneth I.; Steinheiser, Frederick H., Jr.
TITLE	A Bayesian Method for Evaluating Trainee Proficiency. Technical Paper 323.
INSTITUTION	Army Research Inst. for the Behavioral and Social Sciences, Alexandria, Va.
SPONS AGENCY	Office of the Deputy Chief of Staff for Personnel (Army), Washington, D.C.
REPORT NO	ARI-TP-323
PUB DATE	Sep 78
NOTE	52p.
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Bayesian Statistics; *Cutting Scores; Hypothesis Testing; *Mastery Tests; Mathematical Models; *Performance Tests; Predictive Ability (Testing); *Probability; Simulation; *Student Ability
IDENTIFIERS	Test Length

ABSTRACT

A multiparameter, programmable model was developed to examine the interactive influence of certain parameters on the probability of deciding that an examinee had attained a specified degree of mastery. It was applied within the simulated context of performance testing of military trainees. These parameters included: (1) the number of assumed mastery states--master, nonmaster, and perhaps intermediate (likely to soon achieve mastery); (2) the prior distribution of scores from similar examinee groups; and (3) the number of test trials or items administered. The results of several simulations showed that the degree of confidence that a decisionmaker can have about the testee's mastery is markedly affected by the values for the three parameters, and the effects of their combination. Using the Bayesian model, test length and costs could be reduced--as long as the prior information was accurate and valid for the particular group of examinees. Results of the simulation also showed that a test may be too short to be of decision-making value. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED169096

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Technical Paper 323

A BAYESIAN METHOD FOR EVALUATING TRAINEE PROFICIENCY

Kenneth I. Epstein

and

Frederick H. Steinheiser, Jr.

UNIT TRAINING & EVALUATION SYSTEMS TECHNICAL AREA

U. S. Army

Research Institute for the Behavioral and Social Sciences

September 1978

Approved for public release; distribution unlimited.

IM008 492

U. S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the
Deputy Chief of Staff for Personnel

JOSEPH ZEIDNER
Technical Director

WILLIAM L. HAUSER
Colonel, US Army
Commander

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U. S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-P, 5001 Eisenhower Avenue, Alexandria, Virginia 22333

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U. S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Paper 323
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): A BAYESIAN METHOD FOR EVALUATING TRAINEE PROFICIENCY
5. TYPE OF REPORT & PERIOD COVERED: --
6. PERFORMING ORG. REPORT NUMBER: --

7. AUTHOR(s): Kenneth I. Epstein and Frederick Steinheiser, Jr.
8. CONTRACT OR GRANT NUMBER(s): --

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Avenue, Alexandria, VA 22333
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 2Q762722A764

11. CONTROLLING OFFICE NAME AND ADDRESS: Army Deputy Chief of Staff for Personnel, Washington, DC 20310
12. REPORT DATE: September 1978
13. NUMBER OF PAGES: 32

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): --
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: --

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

--

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Decisionmaking
Bayes' Theorem
Mastery
Classification error

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In any testing or evaluation program, there will be some percentage of false positives and false negatives, i.e., misclassifications will occur. A decisionmaker therefore needs to make a best estimate about the true level of proficiency of an examinee. A multiparameter, programmable model was developed to examine the interactive influence of certain parameters on the probability of deciding that an examinee had attained a specified degree of mastery through a program of instruction. The parameters, readily obtainable from

(continued)

20.

decisionmakers, include (a) the number of assumed mastery states ("master," "intermediate," "nonmaster"), (b) the prior distribution of scores from similar examinee groups, and (c) the number of test trials or items that could be given.

Results of several simulations showed that the degree of confidence that a decisionmaker can have in his decision (e.g., "x%" certainty that an examinee is a master) is markedly affected by values for the abovementioned parameters. A key feature of a Bayesian model is that testing time, manpower, expense, and test length can be reduced if the "prior" information is accurate and valid for the particular tested group. If not, little can be gained from a Bayesian model. Simulated test results also showed that a test can be too short to be of any decisionmaking value.

Technical Paper 323

A BAYESIAN METHOD FOR EVALUATING TRAINEE PROFICIENCY

Kenneth I. Epstein

and

Frederick H. Steinheiser, Jr.

UNIT TRAINING & EVALUATION SYSTEMS TECHNICAL AREA

Submitted as complete and technically accurate, by:
Frank J. Harris
Technical Area Chief

Approved By:

A.H. Birnbaum, Acting Director
ORGANIZATIONS AND SYSTEMS RESEARCH LABORATORY

Joseph Zeidner
TECHNICAL DIRECTOR

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333

Office, Deputy Chief of Staff for Personnel
Department of the Army

September 1978

Army Project Number
2Q762722A764

Unit Training Standards and Evaluation

Approved for public release; distribution unlimited.

ARI Research Reports and Technical Papers are intended for sponsors of R&D tasks and other research and military agencies. Any findings ready for implementation at the time of publication are presented in the latter part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

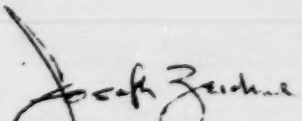
FOREWORD

The research presented in this report was conducted under Project METTEST (Methodological Issues in Criterion-Referenced Testing), under the auspices of the Unit Training and Evaluation Systems (UTES) Technical Area of the Army Research Institute for the Behavioral and Social Sciences (ARI). The goal of Project METTEST is to provide quantitative methods for evaluating unit proficiency. The means for achieving this goal include basic research in test construction methodology, measurement and scaling models, and decisionmaking implications of test score interpretation. ARI Technical Paper 306 is the initial publication on the project.

Related, ongoing programs within the UTES Technical Area include evaluation of small combat units under simulated battlefield conditions (REALTRAIN), qualification of tank gunnery crews and revision of Table VIII (IDOC), and improving the standardization and reliability of the Army Training and Evaluation Program (ARTEP).

Anticipated future research under Project METTEST includes the development of a computer-programed model for performance evaluation and several additional 6.1 basic research grants for the development of measurement, scaling, scoring, decisionmaking, and quality control models for use in performance evaluations when criterion-referenced testing procedures are employed.

The present research was conducted by personnel of the UTES Technical Area as an in-house research project, under Army Project 2Q762722A764. G. Gary Boycan supplied a key creative insight into the "misclassification problem." An earlier version of this paper has been printed in the Proceedings of the October 1976 Naval Training Equipment Center (NTEC) Conference.



JOSEPH ZEIDNER
Technical Director

A BAYESIAN METHOD FOR EVALUATING TRAINEE PROFICIENCY

BRIEF

Requirement:

The educational decisionmaker typically wants to know if a student can perform a job at some prespecified level of acceptability. If the student's test score is above the minimal passing standard, the individual may be classified as a master--otherwise, as a nonmaster. The present paper describes a mathematical model that provides maximal classification accuracy with the least number of test items or trials.

Classification Model:

Estimates of several variables must be provided as input to the model, which is derived from Bayes' Theorem. Two of these variables are probability estimates: the prior expectation of selecting a master from the student population and the conditional probability that a known master would answer a randomly selected test item correctly. Two other variables--the minimal passing standard and the number of test items--are under some degree of control by the tester. Furthermore, the effect of the latter two variables is an interaction, because the model shows that classification accuracy is not invariant over different test lengths when the same percent correct score is attained by examinees.

Findings:

A computer simulation of the model demonstrated the effects of simultaneously varying five variables on classification accuracy. The arbitrary nature of defining the criterion for mastery as a percent correct test score was critically evaluated. Testing may be irrelevant in situations where the test length is less than the minimal number of items.

Utilization of Findings:

The model shows explicitly the risks involved in using a given length of test once the tolerance for misclassification error has been specified by the examiner.

A BAYESIAN METHOD FOR EVALUATING TRAINEE PROFICIENCY

CONTENTS

	Page
INTRODUCTION	1
TRAINING TO MASTERY	1
CONSTRUCTION OF THE MODEL	3
Bayes' Theorem	3
Variables of Interest in the Present Simulation	5
Changes in $p(M\|T)$, Assuming Two Mastery States	5
Elaboration to Three Mastery States	11
Flow-Chart Analysis of How the Bayesian Model Was Developed	18
TEST LENGTH AND MISCLASSIFICATION ERROR	21
SUMMARY AND CONCLUSIONS	25
APPENDIX. A COMPUTATIONAL EXAMPLE FOR THREE MASTERY STATES	27
DISTRIBUTION	31

LIST OF FIGURES

Figure 1.	Conditional probability of mastery when there are two very distinct prior states of mastery	7
2.	Conditional probability of mastery when there are two relatively distinct prior states of mastery	8
3.	Conditional probability of mastery when there are two equally probable prior states of mastery	9
4.	Conditional probability of mastery when there are three prior states of mastery and three conditional probabilities of answering an item correctly	12
5.	Conditional probability of mastery when there are three prior states of mastery and three conditional probabilities (same values in different order from Figure 4) of answering an item correctly	13

Page

Figure 6. Conditional probability of mastery when there are three prior states of mastery (with mastery and nonmastery being the least probable) and three conditional probabilities of answering an item correctly 14

7. Conditional probability of mastery when there are three prior states of mastery (values from Figure 6) and three conditional probabilities of answering an item correctly (values from Figure 5) 15

8. Conditional probability of mastery as a function of percent correct using the same parameter values as in Figure 1A 22

9. Conditional probability of mastery as a function of percent correct using the same parameter values as in Figure 1D 23

Flow Chart 1. Required inputs and sequence of steps in order to obtain the conditional probability of mastery given a test score 19

A BAYESIAN METHOD FOR EVALUATING TRAINEE PROFICIENCY

INTRODUCTION

No instructional system is complete without a strong testing component. Any student who begins an instructional program should be able to achieve all the objectives that the program was designed to teach. However, some students may require remedial or other supplementary instruction to master all of the objectives, even though the program was carefully developed. Furthermore, during the development of the instruction, test data from prospective students are required, first to revise and later to validate the instruction. To support the instructional development activities and to make decisions about the abilities of students who have completed instruction, a powerful testing program is necessary.

The final desired output of a test for a given examinee is information that can pinpoint ability to do whatever is required by an objective. That is, the examiner observes a test score and then infers the ability of the examinee. This paper outlines a "Bayesian" method for drawing such inferences. It also discusses and illustrates the adequacy of the method as a function of the number of test items administered and the effects of the tester's beliefs about the quality of the examinee population on the inferences drawn.

Using the Bayesian method, the testers hypothesized varying numbers of ability groups so that the classification of examinees into these ability groups is most useful to the overall instructional system. For example, the simplest case is to classify examinees into two groups, the first group containing those who have mastered the objective, and the second containing those who have not. Alternatively, one could hypothesize three groups, consisting of masters, nonmasters, and an intermediate group containing people whose skills are almost satisfactory and who could be brought up to the mastery level with relatively little additional instruction. The Bayesian model presented in this paper explores up to three levels of mastery, although this number could easily be expanded. The model also explores the effects on decisionmaking (correctly classifying masters and nonmasters) if more than two ability levels have been hypothesized but are then collapsed to form just two groups--masters and nonmasters.

TRAINING TO MASTERY

Ideally, the educational decisionmaker wants to know if a person (student, trainee) can do a job at some prespecified level of acceptability. A student who scores above the minimal passing standard on a test may be classified as a master; if the score is below the minimal

passing score, the student would be termed a nonmaster. But since data always have some error variability, misclassifications are likely to occur.

		True competency state: Master	True competency state: Nonmaster
Classification based on test score	Master	True positive	False positive
	Nonmaster	False negative	True negative

Ideally, the probability of a true positive should be much greater than that for a false positive, and the probability for a true negative should be much greater than that for a false negative.

To evaluate how well our testing program achieves this goal, we want to be able to infer as accurately as possible the conditional probability of the mastery (or nonmastery) state, given the test score data, $p(M1|T)$, $p(M2|T)$. Our first problem is what amount of data is this probabilistic inference based upon? Suppose that the passing standard was 80% of the test items correct. A student with 33 out of 40 items correct would pass and would be classified as a master. Now suppose that on another form of the test (or a test given over the same material by another instructor), another student gets 25 out of 30 test items correct. This student would also have met the 80% correct criterion and would be classified as a master. The model presented in this paper will show that the $p(M1|T)$ varies systematically with the number of test items, along with the minimal percentage correct for passing.

We may also ask: How is the accuracy of inference about mastery affected by postulating more than two states (mastery and nonmastery)? and can the data from various states be combined without seriously affecting the final $p(M1|T)$ inference? For example, suppose that there are intermediate states of partial mastery. The following decision model shows that $p(M1|T)$ can be more validly estimated when the mastery states are processed independently, but that educational decisionmakers will not sacrifice very much classification accuracy if indeed they do dichotomize multichotomous data. We suggested that defining an intermediate group which required minimal remediation might be useful for some instructional systems. The model shows that the probability of being in the mastery group when indeed the datum was a test score

obtained by a master will be increased if the other data are processed independently. The concept of "independent processing" requires that all nonmastery groups maintain their integrity, rather than being aggregated into one generalized nonmastery group.

CONSTRUCTION OF THE MODEL

Bayes' Theorem

The statistical model which we have applied for classifying students into mastery and nonmastery groups, given their test scores, is based upon a form of Bayes' Theorem:

$$p(M1|T) = \frac{p(T|M1)p(M1)}{[p(T|M1)p(M1) + p(T|M2)p(M2)]}$$

Here we assumed that the two states of nature (master and nonmaster) are mutually exclusive and collectively exhaustive, and that T is the test score observed. We also assume that the test is dichotomously scored and that the items are independent. A correct response is denoted "1," an incorrect response is denoted "0," and the total test score is simply the number of correct responses. What we seek to find is the term on the left, the probability that a given student is a master, having been given his test score. To find it, we need an estimate of the prior probability of mastery ($p(M1)$) in the population of students from which this student was drawn. The prior probability of mastery can be considered the proportion of students in the examinee population we think are masters. For example, if our instruction were very good, the prior probability of mastery would be high, and most of the students who completed the instruction should have mastered the objective. The actual number specified for the prior probability of mastery may be an informed guess based on experience, or it may be based on the empirical results of tests given to previous classes of similar students.

We must also estimate the conditional probability of a certain test score, given that the student who receives that score is a master. For example, if only one item is administered, the conditional probability of a score of one correct, given that the student was a master, is simply the probability that a master responds correctly. We may estimate this conditional probability empirically based on previous student groups, or we may provide a best guess as to how well masters perform, or this conditional probability may reflect a minimal standard of achievement. We shall show how the $p(M|T)$ will vary as a function of the prior expectations of the tester, number of test items, and conditional probabilities, $p(T|M)$, after an example to illustrate the computations.

Suppose that a student chosen at random from a trainee population is given a criterion-reference test, and that he passes the test. Given the results of the test, what is the probability that the student is indeed a master of that particular course of instruction? To calculate the probability, we obtain the following information from the educational expert who administered the CRT: The probability that a master would obtain a passing score = .90, ($p(T|M1) = .90$); the probability that a nonmaster would obtain a passing score = .05, ($p(T|M2) = .05$); and the prior probability of randomly selecting a master from this trainee population is equal to .70, that is, we believe that 70% of this and similar previous trainee populations may be assumed to be composed of masters. Substituting these values into the formula

$$p(M1|T) = \frac{.9 \times .7}{.9 \times .7 + .05 \times .3},$$

equals .977. Hence, before the test score was available, the probability that this student was a master was .70, but after a passing score was observed, the probability that this person is a master has increased to .977. (The probability of this student's being a nonmaster, given the same passing score, $p(M2|T)$, would be equal to 1 - .977 or .023.)

To generalize the Bayesian approach to a wide variety of applications in evaluating training effectiveness, two additions must be made to the basic formula. These additions are the number of trials or items on the test (N), and the number of hypothesized mastery states (S). The derivation of the general Bayesian formula for this purpose was originally presented by Hershman[1]:

$$p(Mi|T) = \frac{\prod_{j=1}^{N} p(Mi|t_j)}{p(Mi)^{N-1} \sum_{i=1}^{S} \frac{\prod_{j=1}^{N} p(Mi|t_j)}{p(Mi)^{N-1}}}$$

In this formula, $p(Mi|t_j)$ equals the conditional probability of a person in the ith mastery state getting the jth test item correct; $p(Mi)$ is the prior probability of the representation of the ith mastery state in the student population (the percentage of students who are estimated

---

[1] Hershman, R. L. A Rule for the Integration of Bayesian Opinions. Human Factors, 1971, 13, 255-259.

to be in the ith mastery state); and p(Mi|T) is the conditional probability of a particular student being in the ith mastery state given his total test score. A computational example showing how the formula is applied for three mastery states is given in the appendix.

Variables of Interest in the Present Simulation

In the typical situation for evaluating training proficiency, the tester has some control over the number of items or trials that he will include on a test. In a performance-based test, each trial may be rather expensive (such as tank gunnery or field artillery, where each shell costs over $100), and so the tester will be obliged to use a minimum number of trials to meet his decisionmaking requirements. Consequently, we examined the effect on p(M|T) when N took on values of 5, 10, 20, and 40 trials.

The tester also has responsibility for assigning reasonable values to the prior probabilities of mastery, denoted as p(Mi), and to the conditional probabilities of a known master (or nonmaster) getting a randomly selected item correct, denoted as p(t|Mi). Values for both the prior and conditional probabilities were systematically manipulated in the present simulation.

The number of mastery states is a variable which the trainer and/or tester may also set. In some measurements of trainee proficiency it may be most appropriate to dichotomize on an all-or-none basis, whereas other training evaluation contexts may suggest a "pass, give refresher training, recycle failures through complete training" trichotomy. More than three mastery states may of course be hypothesized, but the computations in the present and all other models of proficiency evaluation become extremely complex. (However, we are developing a computer program that will handle up to five states of mastery.)

The dependent variable of main interest is the percent of items answered correctly. The tester may decide that 70% is a passing score. But the 70% value is not an absolute standard, since it is dependent upon the number of test items and the prior and conditional probability estimates. In the present simulation, three values of percent correct observed scores were used: 60%, 70%, and 80%.

Changes in p(M|T), Assuming Two Mastery States

The fundamental purpose of the present study was to investigate how the probability of mastery classification changes as a function of the simultaneous manipulation of up to four parameters (independent variables). The scope of the study is not exhaustive, since only several values of each of the four variables were used. However, some general trends do seem to emerge, as can be seen in the following figures.

Figures 1, 2, and 3 show the results of applying the model to a situation in which only two mastery groups (mastery and nonmastery) have been hypothesized. The data points represent the probability that a trainee is a master, given (conditional upon) his total test score, $P(M|T)$. The lines show how the $P(M|T)$ changes as a function of variations in the four parameters: prior expectation of mastery, the percentage correct items observed, the conditional probabilities of both a master and a nonmaster responding correctly to an item, and the number of items comprising the test.

Figure 1 represents a testing situation in which the training was of extremely high quality, since the proportion of masters in the trainee population was assumed to equal 0.9. That is, $p(M1) = 0.9$. Figure 1A portrays the situation in which both masters and nonmasters have attained a rather high degree of proficiency, since the probability of a master responding correctly to any given item is 0.9, and the probability of a nonmaster responding correctly is 0.6. If a person scores 80% on a 5-item test, the probability that he is a master is approximately .91. This probability drops to .65 if a 60% score on 5 items (3 out of 5 correct) is obtained. Note that when the test length is increased to 40 items, an 80% score (32 correct) produces a .99 probability of mastery. However, a score of 60% (24 correct) yields an essentially zero probability of mastery. The effect of the test length variable on classification accuracy is dramatic: If the $p(M|T)$ had to be at least 0.5 for a person to be called a master, then scores of 60% on a 5-item test would lead to mastery classification. But a 60% score on a 40-item test would lead to nonmastery classification.

Figure 1A also illustrates the effect of "prior beliefs" on $p(M|T)$. One might suppose intuitively that the chances were much higher that a person who obtained a score of 60% (even from a 5-item test) came from a population whose probability of correctly answering an item was 0.6 than from a population whose probability of answering an item correctly was 0.9. However, the relative proportions of the two groups (expressed as prior belief in mastery and nonmastery, or $p(M1) = .9$ and $p(M2) = .1$, respectively) are such that the probability of a person being in the mastery state is approximately 0.65 for a score of 3 correct (60%) on a 5-item test. Only by increasing the number of test items can the strong prior bias in favor of the mastery decision be reversed. Figures 2A and 3A show what happens when prior beliefs are not so heavily biased in favor of mastery. In neither case is the probability of being in the mastery state above 0.5 for scores of less than 80%. But Figure 1A suggests that when prior beliefs heavily favor one group over the other, longer length tests should be used. Otherwise, the amount of data may not be sufficient to force a change in the originally held prior beliefs.

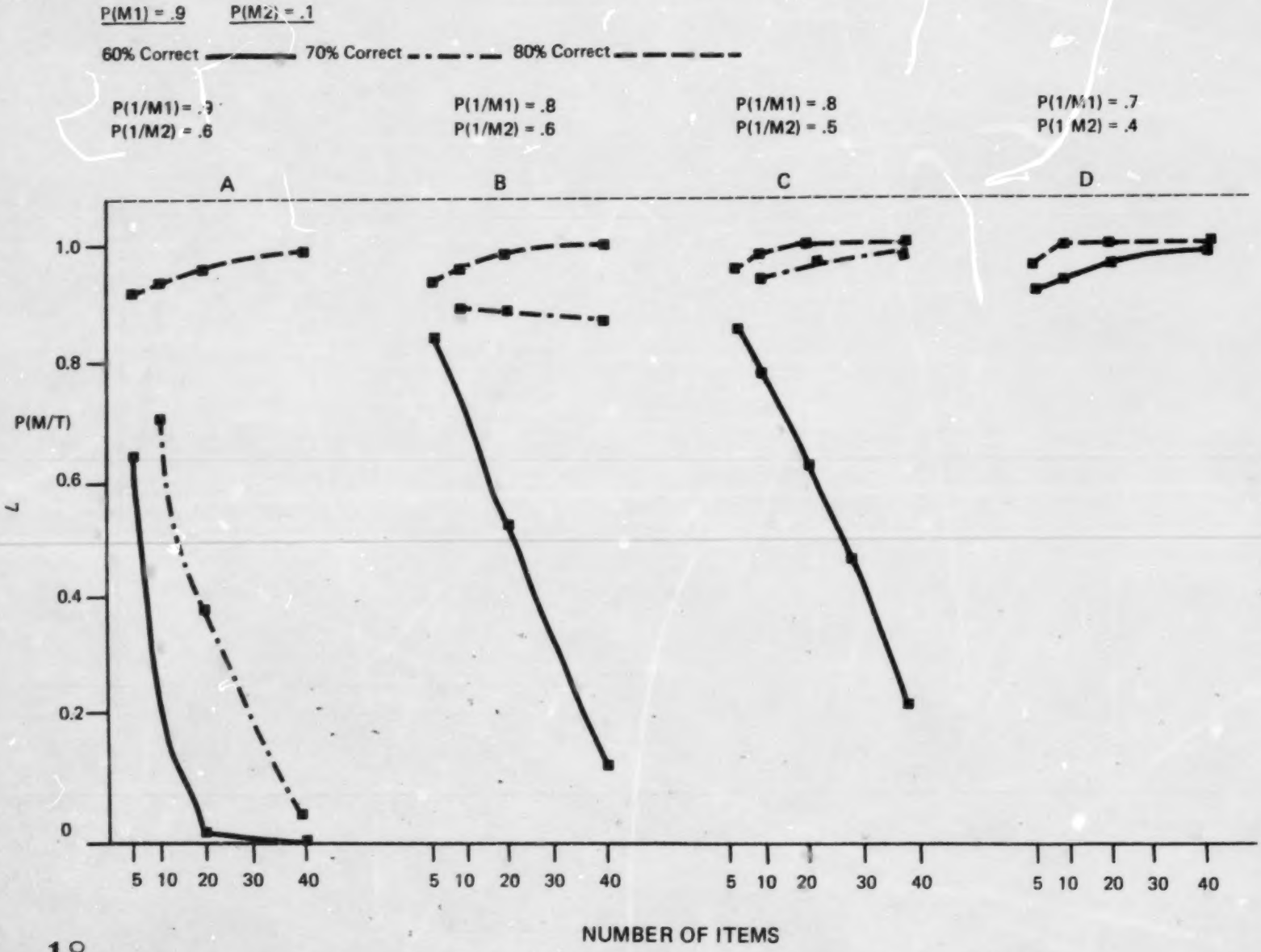


Figure 1. Conditional probability of mastery when there are two very distinct prior states of mastery

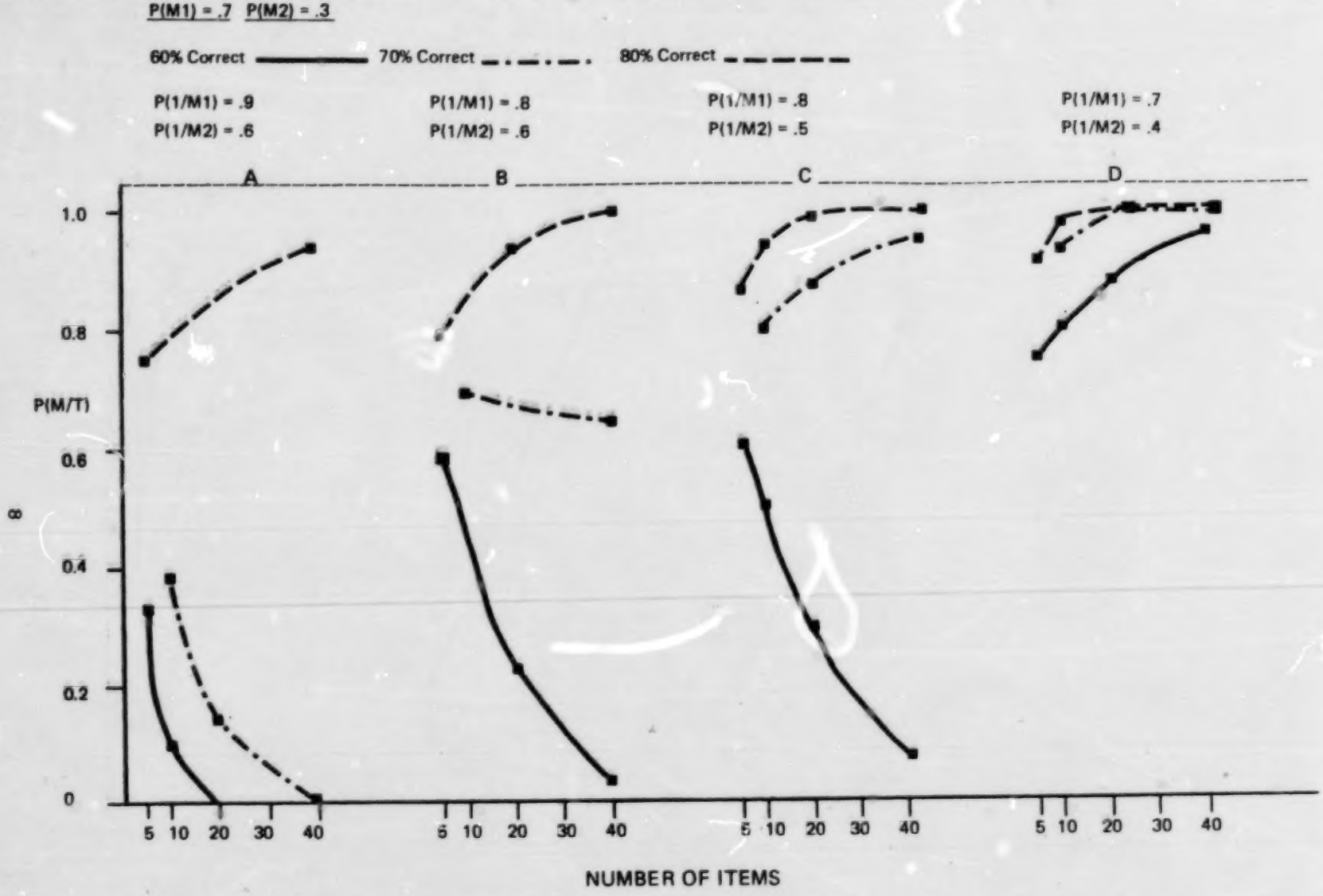


Figure 2. Conditional probability of mastery when there are two relatively distinct prior states of mastery.

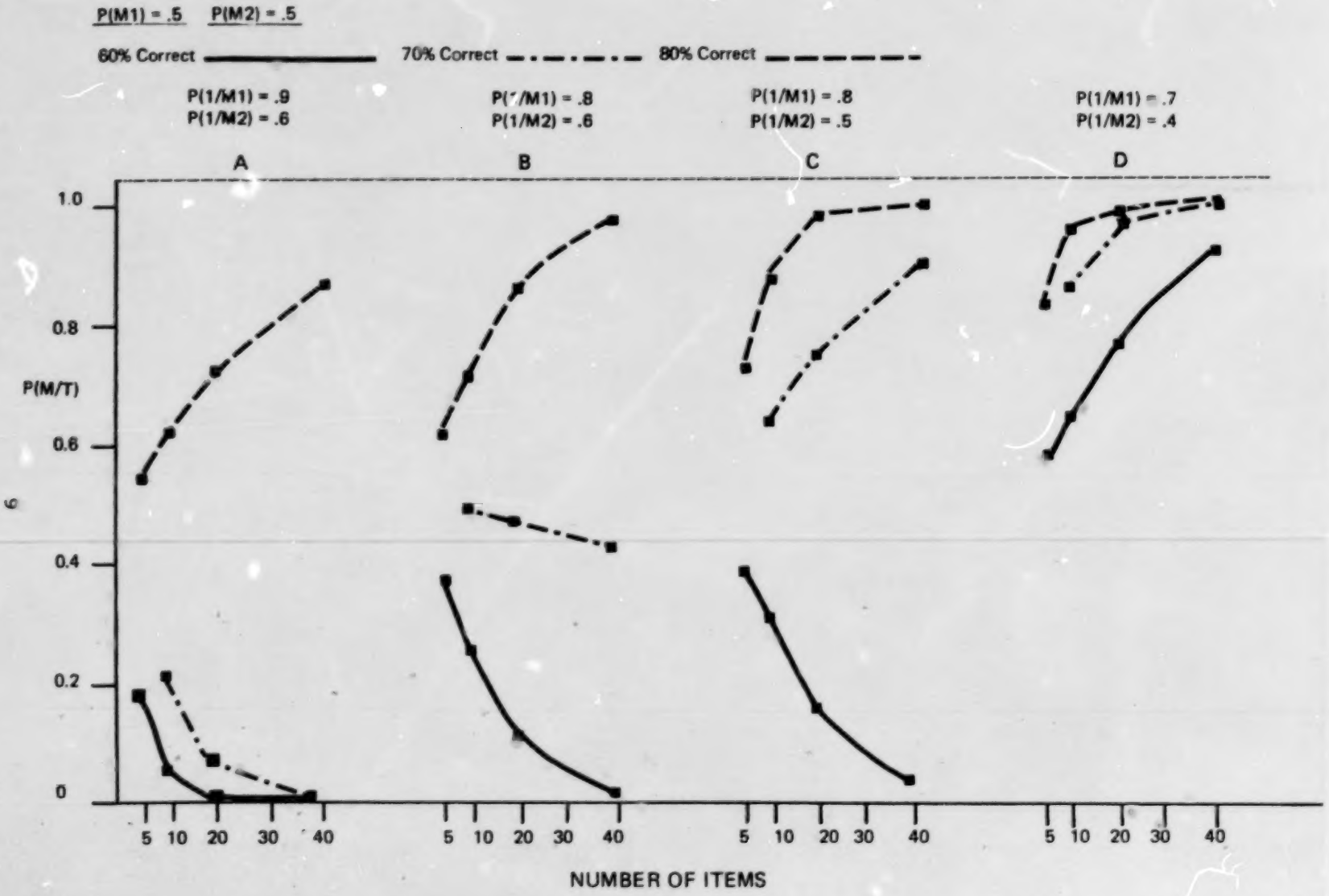


Figure 3. Conditional probability of mastery when there are two equally probable prior states of mastery.

The effect of changing the prior beliefs concerning the proportion of masters and nonmasters in the examinee population, while holding all other parameters constant, can be seen by comparing corresponding Graphs A, B, C, and D in Figures 1, 2, and 3.

The impact of prior information on classification accuracy is very significant: positively so, if the priors are accurate; and unfavorably, if the priors are inaccurate. Novick and Lewis[2] claim that if the criterion level for mastery is kept constant, then low priors will require high test scores to convince the (skeptical) decisionmaker that the examinee has attained the criterion level for mastery. Further, high priors will allow lower test scores to convince a (less skeptical) decisionmaker that the examinee had attained the same criterion level for mastery. In summary, if prior information is strong but inaccurate, then longer tests will be needed to overcome this bias; but if the prior information is strong and accurate, then test lengths can be reduced (by 50%, for example) relative to the number of items that would be required to reach the same decision with no prior information.

The effect of changing the probability of a correct response, $p(1|Mi)$, can be seen by comparing Graphs A, B, C, and D for Figures 1, 2, and 3. For example, the only difference between Figure 1A and Figure 1B is that the $p(1|M1)$ changes from 0.9 to 0.8, all other parameters being held constant. (This change might reflect a lower level of required proficiency and, hence, less training, for Graph B than for A. Or perhaps previous test results indicate that masters of the instruction respond to items with a probability of correct response equal to 0.8 rather than 0.9.) In any case, the effect of this small change in the $p(1|M1)$ on the $p(M|T)$ is readily apparent. For any test length or observed test score, the probability of being in the mastery state is greater in Graph B than in A. This shift is most obvious for the 70% observed correct curve. Notice that $p(M|T)$ on Graph A for an observed score of 70% (28 out of 40 correct) is approximately 0.04. However, the value for $p(M|T)$ in Graph B for 70% of a 40-item test correct is 0.87.

The main reason for this abrupt change from Graph A to B (in Figures 1, 2, and 3) is the lowered requirement for mastery, from 0.9 to 0.8. The probability that "0.9 persons" score only 70% correct on long tests is relatively low. But when masters are defined as those trainees who come from a population with a probability of responding correctly equal to 0.8, the probability of their scoring 70% on a long test is high. One of the most difficult jobs for an instructional designer is

---

[2] Novick, M. R., & Lewis, C. Prescribing Test Length for Criterion-Referenced Measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Center for the Study of Evaluation Monograph Series in Evaluation, III: Problems in Criterion-Referenced Measurement. Los Angeles: U.C.L.A. Center for the Study of Evaluation, 1974.

to describe the level of capability required of graduates and the level of capability actually achieved. Comparison of these graphs indicates the magnitude of the effect that these specifications can have on the classification of trainees.

Graphs C and D of Figures 1, 2, and 3 further illustrate the effect of variations in the probability of correct responses. The only difference between Graphs B and C is that the probability of a correct response from a nonmaster decreases from 0.6 to 0.5. The effect of this decrease in correct response probability from a nonmaster is to increase the probability that someone with a score of 70% or 80% will be a master. Note that the 70% and 80% curves are higher in Graph C than in B. Not evident from the graphs is the additional result that nonmasters are less likely to achieve a high score in C than in B, since $p(1|M2) = .6$ in B, and $p(1|M2) = .5$ in C. Finally, Graph D portrays an extreme case in which neither masters nor nonmasters are responding at particularly high levels. However, the level of performance for nonmasters is so low (0.4), that even for observed scores of 60% the probability of being in the mastery state exceeds 0.8 for all test lengths, except for 5 and 10 items in Figure 2, and 5, 10, and 20 items in Figure 3.

Further detailed analysis of these figures is not included in this paper. In comparing the 12 graphs against each other, note the magnitude of the changes in $p(M|T)$ when small changes have been made in the prior beliefs, in the correct response probabilities, and in the percent correct observed responses. The implication is that extreme care must be taken when specifying parameters in a Bayesian approach to testing and decisionmaking. If the parameters are realistic, great savings in testing time and expense, and increased confidence in decisionmaking are possible (Novick & Lewis, 1974). However, if the parameters are not realistic, there is a very real danger of misclassifying many examinees. The next section of this paper deals with an elaboration of the model to three mastery states, thus helping to quantify sources of classification error.

Elaboration to Three Mastery States

Figures 4, 5, 6, and 7 represent cases for which three mastery states have been hypothesized. In Figures 4 and 6 the probability of a correct response for a person assumed to be in mastery state M1 equals 0.8; for mastery state M2 this probability is 0.6; and for mastery state M3, it is 0.5. These values could correspond to the situation in which the nonmastery group was divided in half. That is, those persons whose probability of getting any given item correct is 0.5 (comprising mastery state M3) would need extensive retraining; whereas those whose probability is 0.6 (comprising mastery state M2) would merely need selective retraining. People in mastery state M1 have a probability of 0.8 for making a correct response and may therefore be considered as "masters" who have successfully passed training.

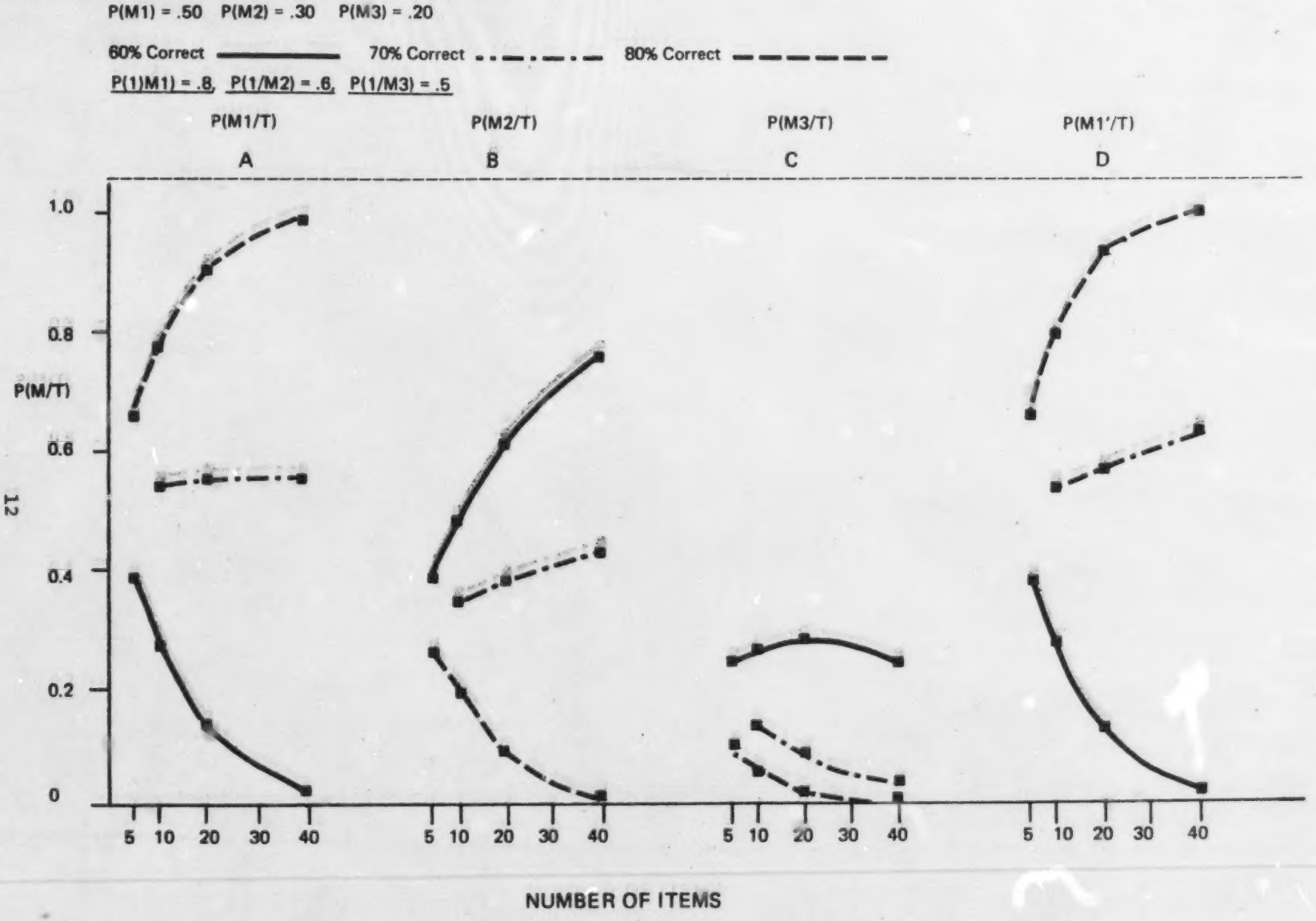


Figure 4. Conditional probability of mastery when there are three prior states of mastery and three conditional probabilities of answering an item correctly.

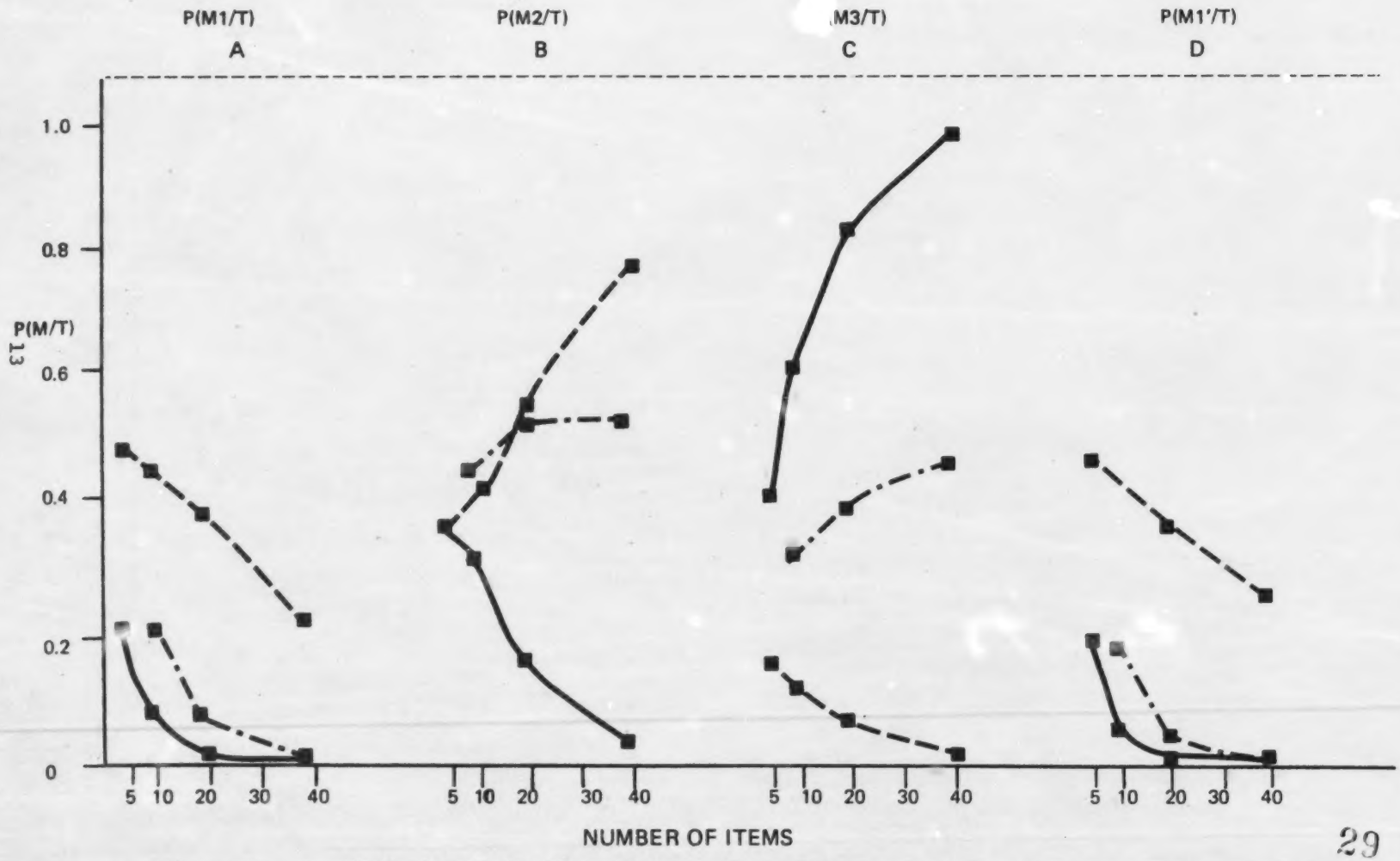


Figure 5. Conditional probability of mastery when there are three prior states of mastery and three conditional probabilities (same values in different order from Figure 4) of answering an item correctly.

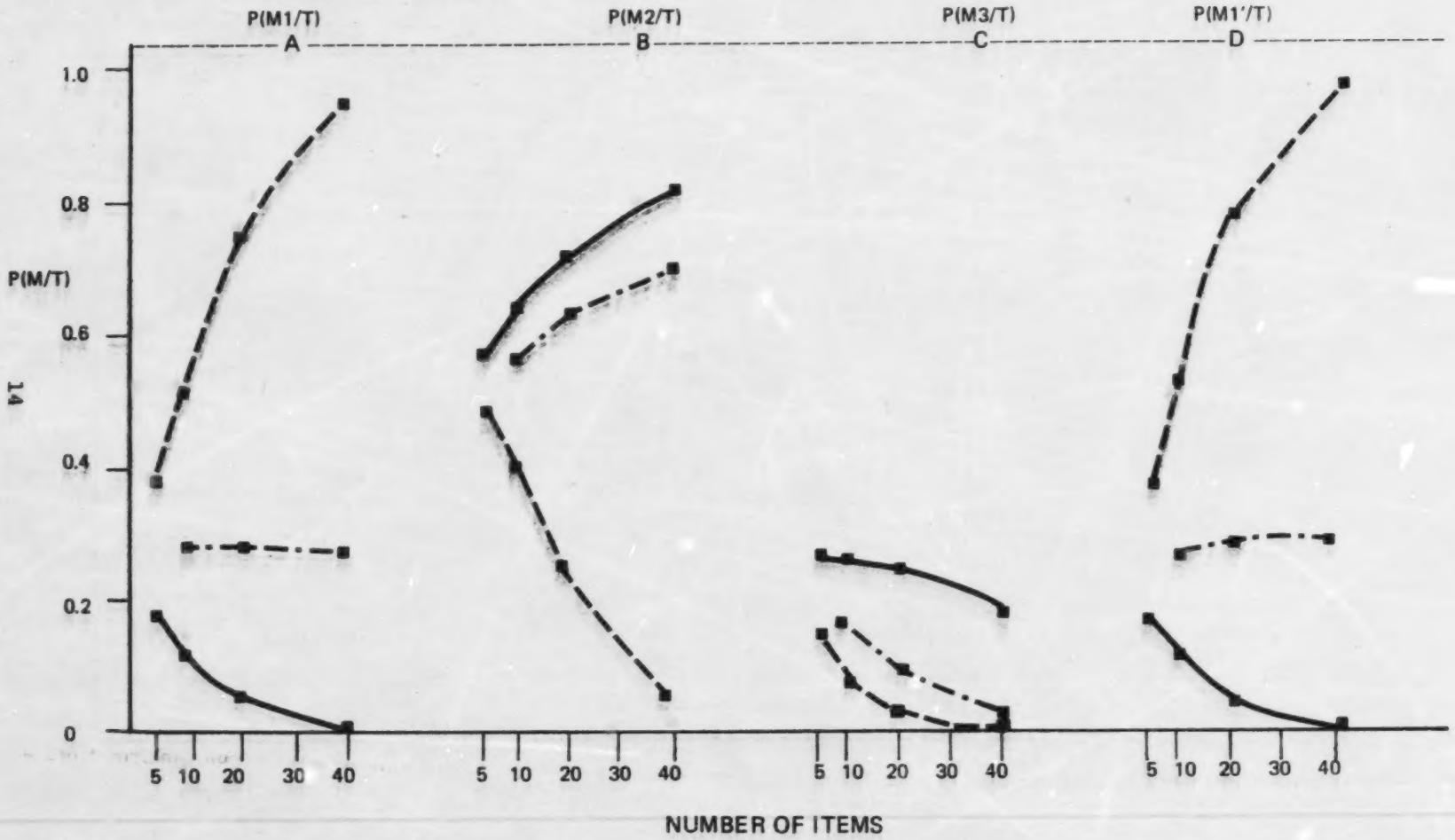


Figure 6. Conditional probability of mastery when there are three prior states of mastery (with mastery and nonmastery being the least probable) and three conditional probabilities of answering an item correctly.

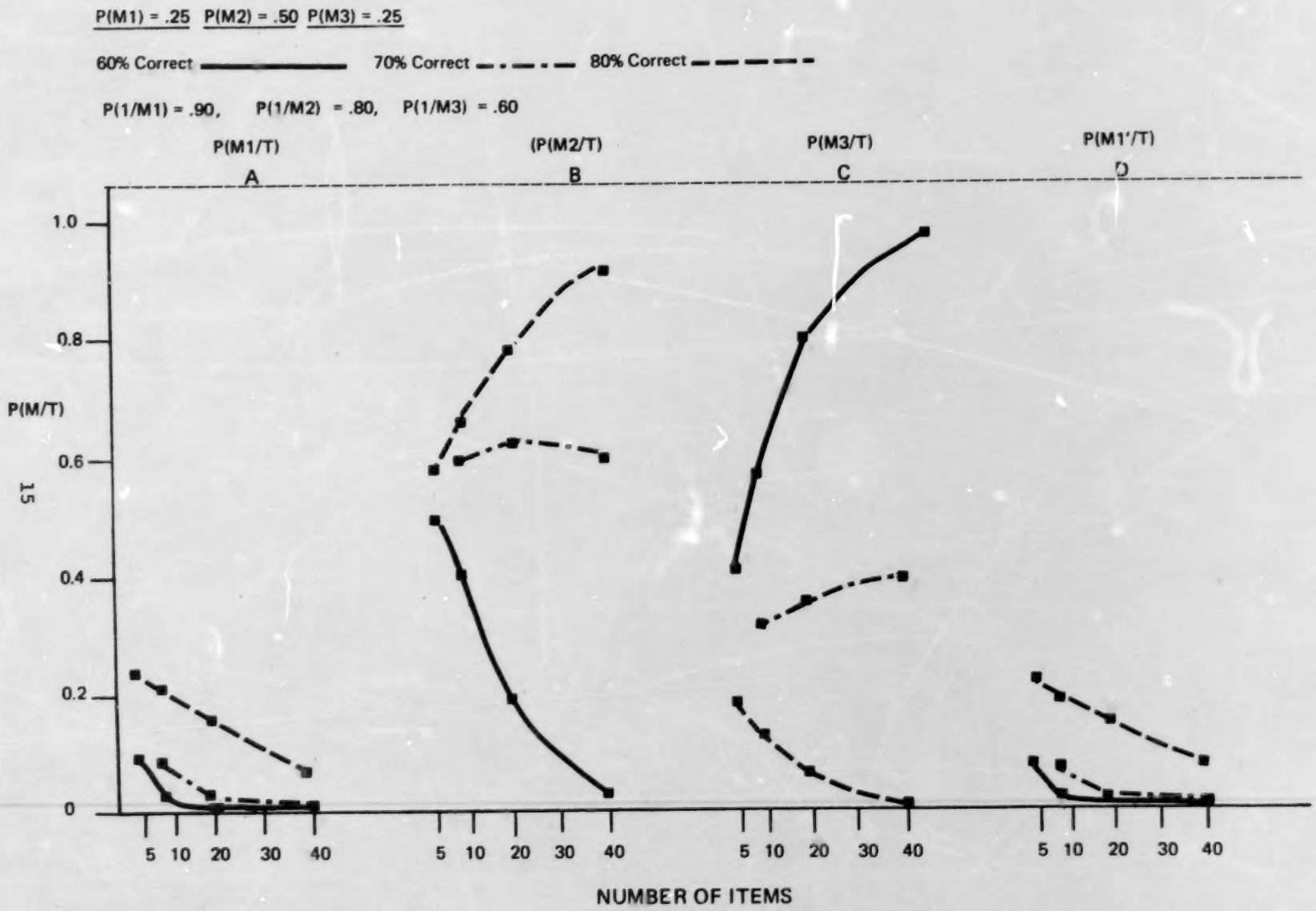


Figure 7. Conditional probability of mastery when there are three prior states of mastery (values from Figure 6) and three conditional probabilities of answering an item correctly (values from Figure 5).

For Figures 5 and 7, the corresponding probabilities of a correct response for people in mastery states M1, M2, and M3 are 0.9, 0.8, and 0.6, respectively. These probabilities might describe a situation in which the mastery group was dichotomized, perhaps in an attempt to identify those students who had achieved an exceptionally high level of proficiency, i.e., $p(1|M1) = 0.9$.

In Figures 4 and 5, the prior probabilities (or assumed proportions) of examinees in each mastery state are: $p(M1) = 0.5$, $p(M2) = 0.3$, and $p(M3) = 0.2$. In Figures 6 and 7, the corresponding prior probabilities are 0.25, 0.50, and 0.25, respectively. The prior values in Figures 4 and 5 display a bias toward higher levels of mastery (50% of the examinees are assumed to be type M1 masters), whereas the bias in Figures 6 and 7 is toward the intermediate level of mastery (50% of the examinees are assumed to be type M2 masters).

A detailed analysis of Figures 4 and 5 provides the basis for an interpretation of Figures 6 and 7, which is an exercise left to the reader. The three graphs labeled A, B, and C represent the probability that an individual is in mastery state M1, M2, and M3, respectively. Graph D represents the probability that a person is in mastery state M1 after mastery states M2 and M3 have been combined into one composite state.

Graph A of Figure 4 shows the probability that an individual is in mastery state M1, given observed scores of 60%, 70%, and 80% correct on 5-, 10-, 20-, and 40-item tests. Thus, for an observed score of 4 out of 5 correct, the probability that this person is in mastery state M1 is about 0.65. But if this same person scores 32 out of 40 (still 80% correct), the probability that he is an M1 master jumps to 0.98. These results are similar to those obtained when two mastery groups were hypothesized, and again illustrate the effect of increasing test length on the level of confidence in the mastery classification $p(M|T)$.

The probability of being in mastery state M2, given observed scores, is plotted in Graph B. If a person got 4 out of 5 correct, the probability of being in state M2 is about 0.25. However, if he got 32 out of 40 correct (still 80% correct), this probability plummets to 0.02. Finally, using these same test score values, Graph C shows that the probability of being a type M3 master is 0.10 for 4 out of 5 correct, and nearly zero for 32 out of 40 correct. This result makes intuitive sense, because there is only 20% of type M3 (non)masters in the examinee population, and the probability of their getting any item correct is only 0.50, which is a long way from 80% observed correct.

Notice that for any given test length and percent correct, the sum of the probabilities of being in states M1, M2, and M3 equals 1.0. Comparison of Graphs A, B, and C shows that when either 70% or 80% of the items for any test length are correctly answered, the probability of being in state M1 is greater than the probability of being in either

state M2 or M3. That is, both the 70% and 80% curves are higher in Graph A than in either Graph B or C. For an observed score of 60%, the probability of being in state M2 is greater than for M1 or M3. The probability of being in state M3 is rather low for all values of test length and percent correct observed in this particular example.

Graph D depicts the probability that a person is in mastery state M1, as opposed to a new nonmastery state composed of both M2 and M3. It can be seen that when states M2 and M3 have been thus combined, the probability of being in state M1 is greater than when all three states were analyzed independently. For observed scores of 70% or 80% correct, there is slight difference in the decisions that would be made under the "independence" versus "composite" conditions. However, if a score of 60% were observed, the possibility of distinguishing between M2 and M3 would be lost when those states were combined. This loss of information may be very important if there is a large difference in cost between the selective training required for people in the M2 state and the extensive retraining needed for those in M3. This example also illustrates the potential significance of maintaining the integrity of the various nonmastery states. If the instructional decisionmaker knew the p(M1) with great accuracy and also knew that there were two nonmastery states, but decided to combine the two states of nonmastery into just one state, he or she would be throwing away potentially valuable information. We shall return to this point in the discussion of Figure 5.

The interrelationship between test length and three hypothesized mastery states becomes even more apparent in Figure 5. For example, Graph A shows that the probability of being in state M1 for 80% correct on a 5-item test is about 0.48. The probability of being in state M2 (shown in Graph B) for 80% correct on a 5-item test is about 0.36. There is thus a greater chance that a person whose score is 4 out of 5 is in M1 ($p(M1|T) = 0.48$), instead of M2 ($p(M2|T) = 0.36$) or M3 ($p(M3|T) = 0.16$). However, if a score of 80% correct were observed on a 40-item test, the graphs indicate that a much different decision would be appropriate. In this case, $p(M1|T)$ equals 0.21, $p(M2|T) = .78$, and $p(M3|T) = 0.01$. Hence, people scoring 32 out of 40 correct should be classified as type M2 masters. Also note that a score of 60% for any test length implies that these people should be placed in the M3 state.

For the data used in Figure 5, the probability of finding M1 type masters is overall quite low. Instead, for the levels of achievement demonstrated by obtained scores of 60%, 70%, or 80%, it is more likely that such scores were produced by people in mastery states M2 ($p(1|M2) = 0.8$) and M3 ($p(1|M3) = 0.6$).

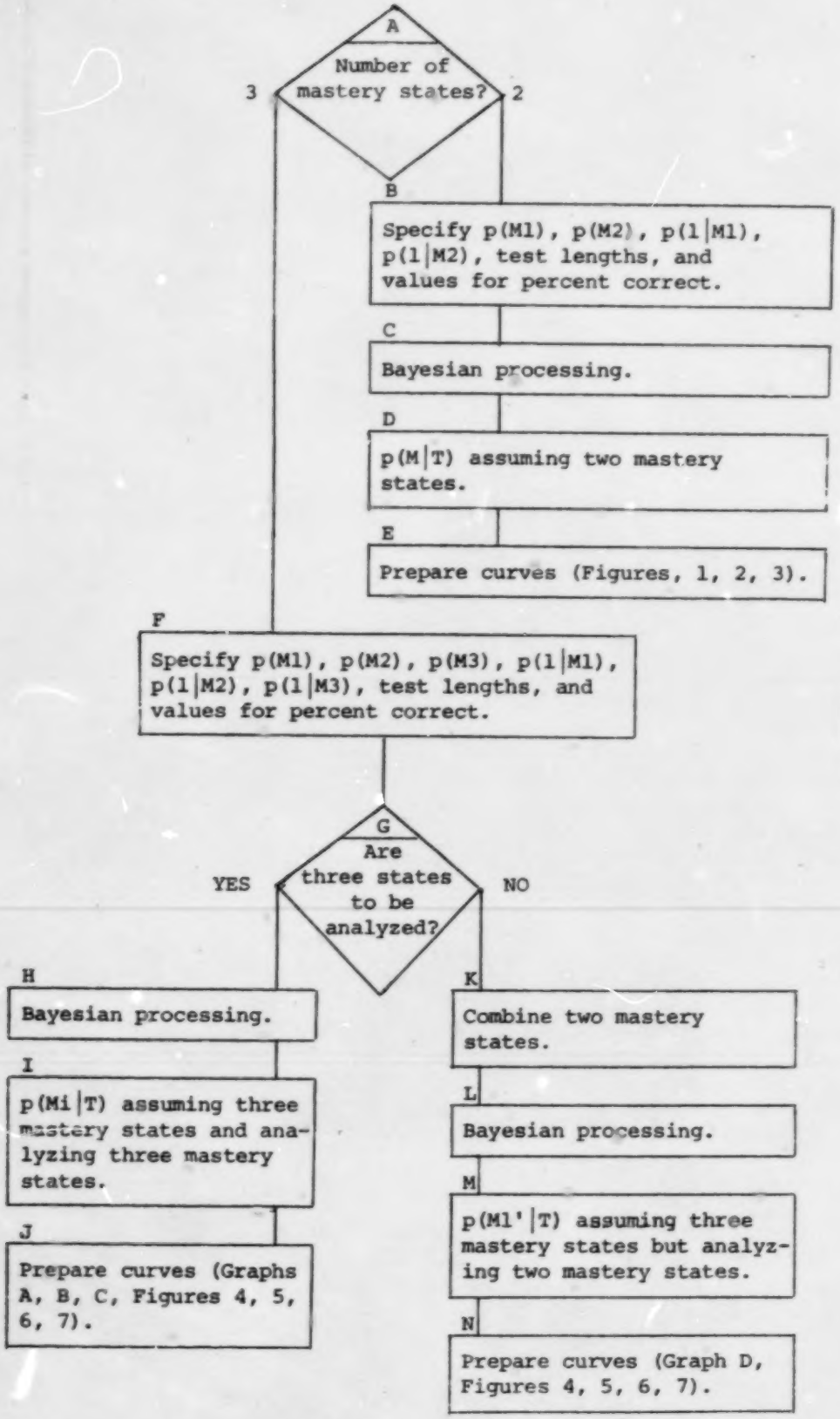
Graph D in this figure also represents the probability that a person is in mastery state M1 as opposed to the new (non)mastery state formed by combining states M2 and M3. In this example, most of the probabilities in Graph D are lower than in Graph A. A glance back at Figure 4, Graphs A and D, reveals that the combination of states M2 and M3 increased the probability of classifying a person with a given test score as a type M1 master. Inspection of the trends in Graphs A and D of Figures 4, 5, 6, and 7 suggests that the effect of combining mastery states is to enhance the trend of the uncombined state. That is, if the probability of being in state M1 is high when the three states are treated independently, the $p(M1|T)$ will increase after M2 and M3 are combined. Conversely, if $p(M1|T)$ is low when the three states maintain their integrity, then combining states M2 and M3 tends to decrease the $p(M1|T)$.

Flow-Chart Analysis of How the Bayesian Model Was Developed

The impact of adding a third mastery state to the development of the model can be illustrated by tracing the logic that is required in formulating a description of the examinee population. (Refer to accompanying flow chart for a schematic summary of this discussion.) The first question the decisionmaker must ask (and which we considered) is: Are there two or three states of mastery inherent in the examinee population (Step A)? If two states are posited, parameter estimates for $p(M1)$, $p(M2)$, $p(1|M1)$, and $p(1|M2)$ are specified, along with plausible test lengths and values for the percent correct (Step B). The output of the Bayesian processing is the probability that a particular person is in the mastery state, $p(M1|T)$ (Step D). A unique graph for each of Figures 1, 2, and 3 was obtained by holding the prior and conditional probabilities constant while simultaneously varying the test lengths and percent correct that would plausibly be observed (Step E). If three states are hypothesized, parameter estimates for $p(M1)$, $p(M2)$, $p(M3)$, $p(1|M1)$, $p(1|M2)$, and $p(1|M3)$ need to be specified, along with values for test lengths and percent correct (Step F).

Now if three states are postulated, a second decision must be made (Step G). It would seem to be usually desirable to determine the probabilities of a person's being in each of the three states (Step I). Having obtained these probabilities for selected values of prior and conditional probabilities and over a range of test lengths and percent correct scores, Graphs A, B, and C can be drawn such as those shown in Figures 4, 5, 6, and 7 (Step J).

However, in some instances it may be more convenient to combine the information known about two of the three mastery states. For example, even though one mastery state and two nonmastery states are hypothesized, the decisionmaking process may require that people be divided into only two groups--"mastery" and "nonmastery." In the present example, states M2 and M3 were combined (Step K). The result


A
Number of mastery states?
3
2
B
Specify p(M1), p(M2), p(1|M1), p(1|M2), test lengths, and values for percent correct.
C
Bayesian processing.
D
p(M|T) assuming two mastery states.
E
Prepare curves (Figures, 1, 2, 3).
F
Specify p(M1), p(M2), p(M3), p(1|M1), p(1|M2), p(1|M3), test lengths, and values for percent correct.
G
Are three states to be analyzed?
YES
NO
H
Bayesian processing.
I
p(Mi|T) assuming three mastery states and analyzing three mastery states.
J
Prepare curves (Graphs A, B, C, Figures 4, 5, 6, 7).
K
Combine two mastery states.
L
Bayesian processing.
M
p(M1'|T) assuming three mastery states but analyzing two mastery states.
N
Prepare curves (Graph D, Figures 4, 5, 6, 7).

Flow Chart 1

of Bayesian processing on these combined data is the probability that a person is in the new mastery state (Step M). Iteration of this procedure for various test lengths and percent correct scores over the same prior and conditional probabilities yields Graph D curves, such as those of Figures 4, 5, 6, and 7 (Step N).

The differences that result from following each of the three paths in the flow chart can be seen by comparing Figures 3A, 5A, and 5D. In each case the prior probability of being in mastery states M1 was set equal to 0.50, and the conditional probability that a type M1 master would make a correct response to an item was set equal to 0.90. Figure 3A corresponds to path A,B,C,D,E in the flow chart. Figure 5A corresponds to path A,F,G,H,I,J; and Figure 5D corresponds to path A,F,G,K,L,M,N.

In Figure 3A, $p(1|M2) = 0.6$, that is, a nonmaster has a 60% chance of correctly responding to an item. However, in Figure 5D the nonmastery state is the combination of states M2 and M3, with probabilities of responding correctly to an item of 0.8 and 0.6, respectively. The effect of combining M2 and M3 is to create a new (non)mastery state, where the probability of a correct response is a weighted average of the values for the uncombined groups. By defining a relatively high ability intermediate state and then combining it with a relatively low state, the probability of being in the highest mastery state is lower than if that intermediate state remained undefined. In fact, if the Figure 5 values of the prior and conditional probabilities are valid representations of the "real" states of mastery, but the values of Figure 3 (which are a simplification of the Figure 5 values) are used for decisionmaking, then people achieving scores of 80% will be falsely classified as type M1 masters.

The differential trend between Graphs A and D of Figure 5 is noteworthy, although the absolute magnitude of the trend is rather small. For different parameter estimates (of prior and conditional probabilities), the effect of combining groups may be much more extensive. Note also that the information provided in Graph D refers only to the probability of a person's being in the mastery state and does not directly show the loss of information about the two discrete nonmastery states that have been combined. Furthermore, when two mastery states are combined and contrasted to a third nonmastery state, the changes in the probability of being in the newly defined mastery state will often be quite different from the probability of being in the original mastery state.

It must be emphasized that unrealistic descriptions of the examinee population (in terms of number of mastery groups) can cause severe distortions in classification accuracy. For example, had the decisionmaker hypothesized only two states when, in fact, training had produced three fairly distinct states of proficiency, the results of his analysis could be highly misleading. Thus, note that the 80% line of Figure 3A

ascends as more items are added (i.e., $p(M1|T)$ increases), whereas the 80% line of Figure 5D descends (i.e., $p(M1|T)$ decreases) as more items are added.

Caution must also be observed in the opposite case, where one might be tempted to specify more states of mastery than are actually present, in an effort to extract more information than is justified by the test data.

The present Bayesian model is not limited to three mastery states. Exploratory analyses have been conducted with up to five mastery states, and it is also hoped that the model can be generalized to deal with continuous distributions.

TEST LENGTH AND MISCLASSIFICATION ERROR

One of the most important questions that must be answered in designing a training evaluation program is "What is the probability of falsely classifying a person on the basis of a given observed score?" It is also possible to turn the question around and ask "How long must a test be, and what score is required for classification decisions to be made with some specified lower limit of misclassification?"

Figures 8 and 9 demonstrate how the Bayesian model can be used to answer these two questions. Assuming that the prior and conditional probabilities are realistic and fixed, the important variables are then test length and cutting score. Suppose that $p(M1) = 0.9$, $p(M2) = 0.1$, $p(1|M1) = 0.9$, and $p(1|M2) = 0.6$ as in Figures 8 and 1A. In this example, the prior belief that an untested trainee is a master is very high, $p(M1) = 0.9$. A reasonable question might therefore be "What score must be observed such that a nonmastery decision can be made with at least 90% confidence?" (In other words, what data are required to force a reversal in the prior belief?)

To be 90% confident of a nonmastery decision, $p(M2|T)$ must be equal to at least 0.90. Since the sum of $p(M1|T)$ and $p(M2|T)$ equals 1.0, $p(M1|T)$ must therefore not be greater than 0.10. Referring to Figure 8, a horizontal line crossing the ordinate at 0.10 can be drawn. This line crosses the curve for a 5-item test at a point corresponding to 26% correct. The next lowest possible test score is one correct (20%), so the decision rule is that all persons scoring one correct or less should be considered nonmasters. The point on the ordinate corresponding to 20% correct on the 5-item test is about 0.05. Hence, the final decision rule states that nonmastery decisions based on an observed score of 1 correct out of 5 can be made with 95% confidence ($1.00 - 0.05 = 0.95$). For observed scores lower than the cutoff score, the confidence in making a correct decision must increase. Continuing with the present example, the $p(M1|T)$ if zero correct are observed is

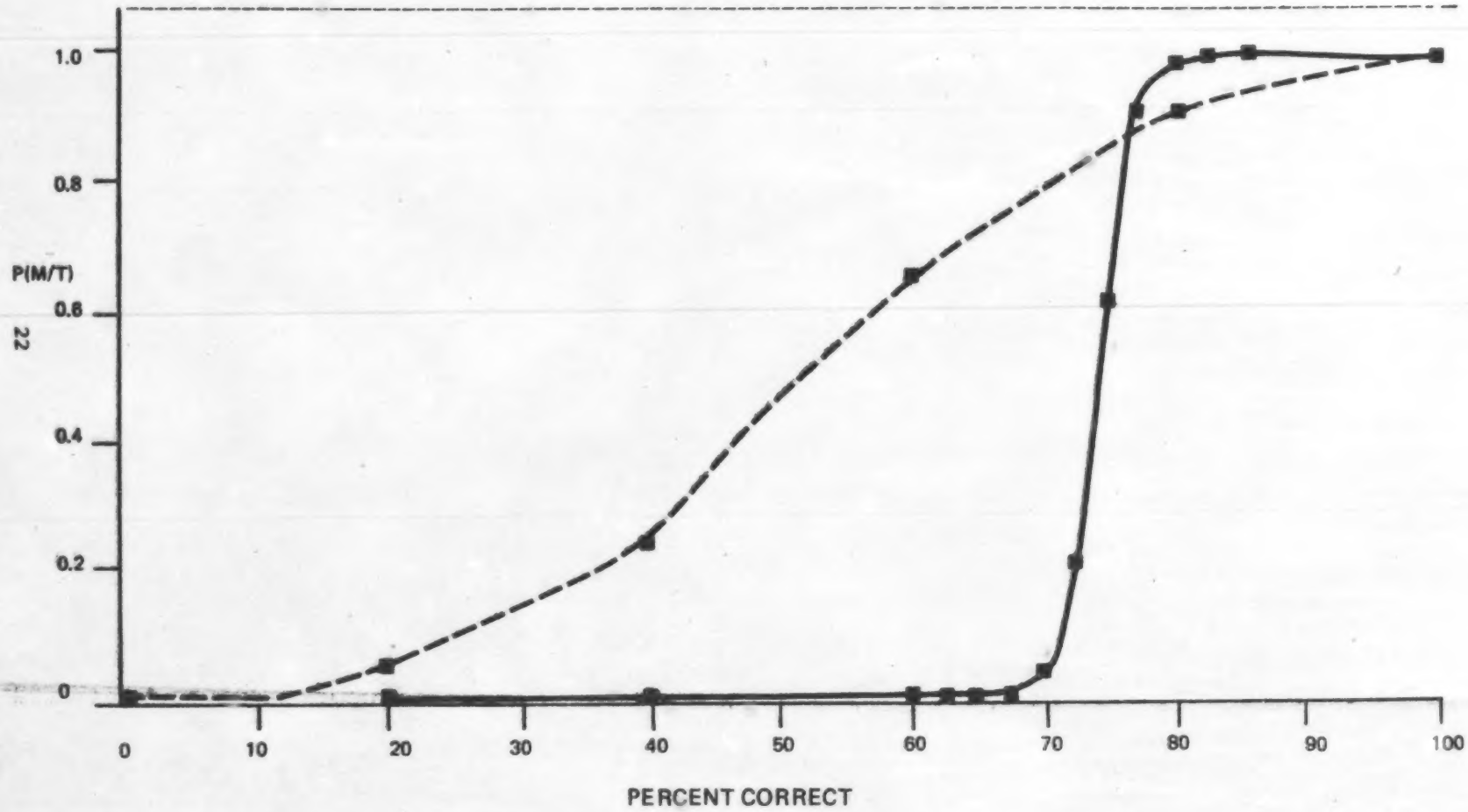


Figure 8. Conditional probability of mastery as a function of percent correct using the same parameter values as in Figure 1A.

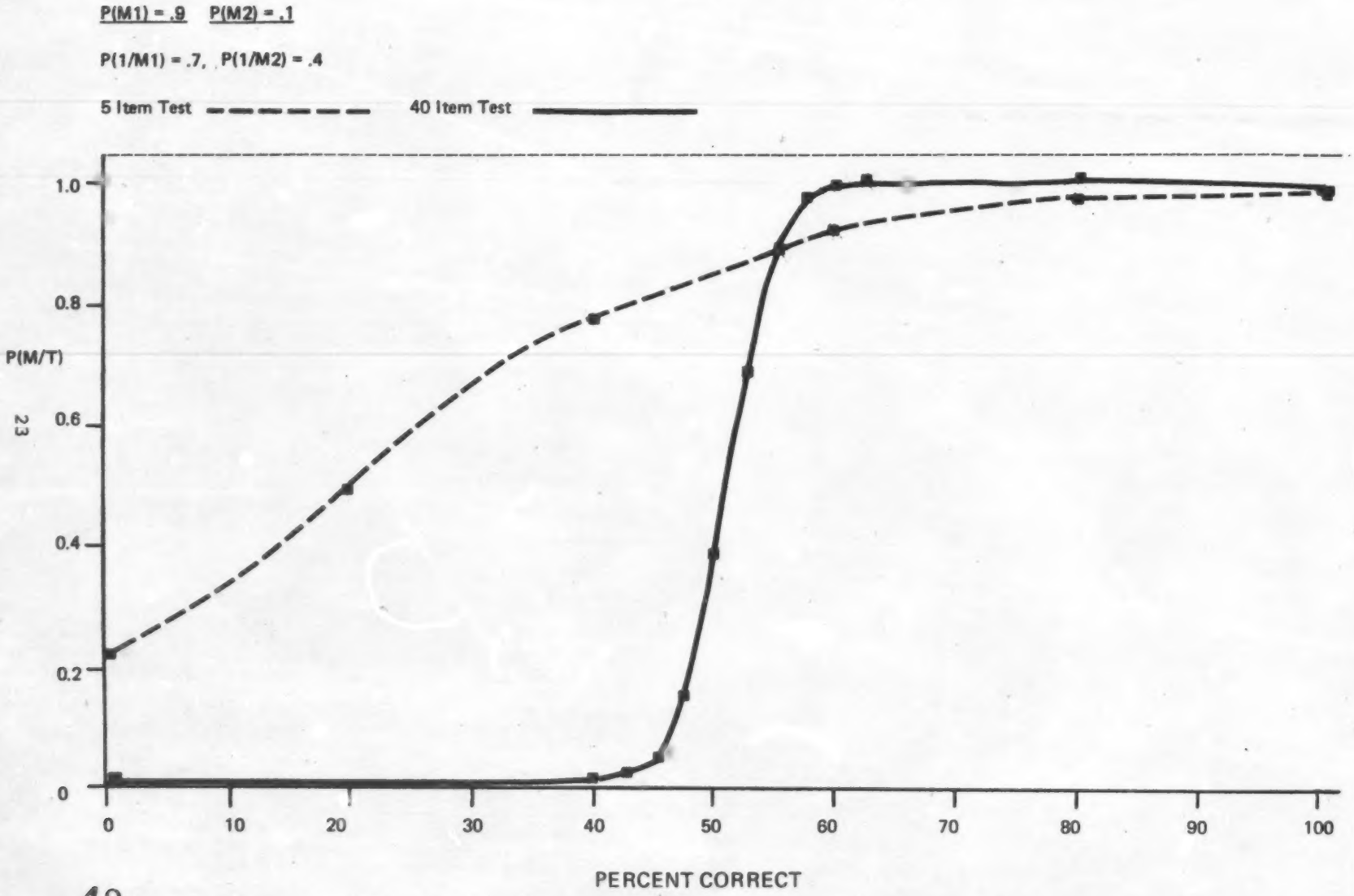


Figure 8. Conditional probability of mastery as a function of percent correct using the same parameter values as in Figure 1D.

virtually equal to zero. Hence, those persons who get no items right may be classified as type M2 nonmasters with nearly 100% confidence.

A similar analysis applied to the 40-item test curve indicates that the cutting score should be about 73% correct. The next lowest possible score to 73% is 28 correct out of 40 items, or 70%. The probability of mastery, given an observed score of 28 correct, is about 0.04. At such a low value of $p(M1|T)$ the chances for misclassification using a 5-item test and a 40-item test are almost the same. However, the observed percent correct at which the nonmastery decision is made for the two tests is 20% on the 5-item test and 70% on the 40-item test. Superficially, two tests of different lengths would seem to produce the same decision outcome, and longer tests may not really be necessary for reducing classification error.

To appreciate the benefits gained from using longer tests, we must examine the entire curve. Note that at 80% correct, the 5-item test yields a $p(M1|T)$ equal to 0.92. This result means that, on the average, 8% of the mastery decisions will be in error, since $p(M2|T)$ equals 0.08. For the 40-item test, the probability of mastery, given 80% correct, is about 0.99. That is, there is only a 1% chance that an examinee of nonmastery competence would be incorrectly classified as a master.

A test that distinguishes sharply between masters and nonmasters is one in which the probability of mastery is close to either 0.0 or 1.00 for most obtained scores. On such tests there is only a small region in which classification error is large. For example, in Figure 8, for the 40-item test the region where $p(M1|T)$ is greater than 0.1 and less than 0.9 extends from 71% to 77% correct. This means that the probability of misclassification (calling a true master a "nonmaster," and vice versa) will exceed 0.10 only when observed scores range from 71% to 77% correct. In contrast, the region of the 5-item test curve for which $p(M1|T)$ is greater than 0.10 and less than 0.90 extends from about 26% to about 79% correct. Hence, there is a much larger region for which the probability of misclassification exceeds 0.10. Therefore, if classification accuracy is to be maximized over the entire range of possible test scores, longer tests are required. Ideally, a very long test would produce a step function, for which the probability of a given mastery state would be very close to either zero or one.

Figure 9 can be analyzed in a manner similar to that for Figure 8. However, Figure 9 has one outstanding characteristic that merits special attention. If nonmastery decisions must be made with 90% confidence, and a horizontal line at $p(M|T) = 0.1$ is drawn, the line does not intersect the curve for the 5-item test. This means that it is not possible to classify a nonmaster with 90% confidence if a 5-item test is used, given the parameters used in Figure 9. If resource or time constraints are such that no more than five items may be given, and if the

parameter values used in Figure 9 are realistic, and if 90% confidence for mastery decisions are required, then there is no reason to test. Testing is irrelevant because no matter what score is observed, including zero correct, the decision rule compels a mastery decision to be made. In fact, for the present values, the probability of mastery, given zero correct, is equal to 0.21. This simply means that if persons obtaining a score of zero are classified as nonmasters, 21% of them will be misclassified, on the average.

The implication of these results for performance testing is obvious. Since performance tests are often rather short, it is essential to recognize the magnitude of misclassification error that can be incurred with such tests. Designing tests that have clear and direct relation to actual performance is certainly a worthwhile and much-needed effort. However, reasonable levels of confidence in classifying trainees must not be sacrificed merely for the sake of using conveniently short tests.

SUMMARY AND CONCLUSIONS

The present simulation study highlights some very pertinent issues for test developers and educational decisionmakers. The simulated results demonstrate explicitly the effects that changes in the estimates of the examinee population quality, number of assumed mastery states, criteria required for mastery classification, and test length can have on the probability of correctly classifying a particular examinee. Furthermore, the simultaneous manipulation of combinations of these parameters can produce drastic and complex changes in the probability of correctly classifying a specific examinee.

A unique feature of any Bayesian model is the need for "prior" information. In the present context, this is the estimate of the proportion of masters and nonmasters in the examinee population. The more accurately that such an estimate can be made, the greater the value in using a Bayesian approach: "It is this increment in information that is equivalent to prior observations which permits a reduction in test length when a Bayesian procedure is used" (Novick & Lewis, 1974, p. 149, italics added). If the number of items or trials that can be given on a test is constrained (such as the cost associated with firing live ammunition in tank gunnery or field artillery), then a Bayesian model may be desirable.

The simulation results also demonstrate that a criterion for mastery (usually expressed as a percent correct of all possible test items that could be given) is not invariant across various test lengths. The significant implication is that the probability of correct classification varies as a function of test length, mastery criterion, and their interaction. Classification accuracy improves with longer length tests and with stricter mastery criteria. However, there is a point of

diminishing returns, for which increases in test length or criterion strictness yield successively smaller increments in classification accuracy.

Another unique feature of the Bayesian approach is that it yields the probability of a mastery state, given or conditional, upon a specific examinee's test score. Since the mastery state is probabilistically inferred and not assumed, it is not possible to compute false positive and false negative error rates. However, the model seems to be asking the correct question: "What is the probability that a given examinee is a master, given his test score?" An alternative binomial model does give the false positive and false negative error rates but does not give explicit information about a specific examinee. This is because it assumes a certain mastery state and then works "backwards" to complete the misclassification rates for that hypothesized mastery state, instead of using prior data to infer the unobservable mastery state.

Hershman's (1971) original formulation of the Bayesian model combined several states of nature into a smaller number of states, under the assumption that the prior probabilities of the new states were equal. This assumption leads to the conclusion that it is generally undesirable to combine states of nature (mastery) because of the severe distortions in classification accuracy that arise. In contrast, our approach was to simply combine the prior probabilities, but not to equate them as Hershman did. Hence, $p(M1) = .25$, $p(M2)$, $= .3$, and $p(M3) = .45$ would be combined into the values $p(M1) = .25$ and $p(M2,3) = .75$. The effect of this method of combining prior probabilities caused relatively little change in classification accuracy, compared to the case where the mastery states were processed distinctly. Our approach of combining prior information seems more reasonable, since one would expect that the probability of one state which is not combined with any other should not be affected when the others are combined. This may be called an "independence of states of nature" assumption.

The final rather significant insight to be gleaned concerns the issue of minimal test lengths that are required when limits for the probability of misclassification have been specified by the examiner. It has been analytically shown that a test can be too short to be of any value in decisionmaking, depending upon the misclassification rate that the examiner is willing to tolerate. What this model does is to show explicitly the risks involved in using a given length of test, once the tolerance for misclassification error has been specified by the examiner.

APPENDIX

A COMPUTATIONAL EXAMPLE FOR THREE MASTERY STATES

The following example illustrates the computations necessary for processing data with the Bayesian model. The values chosen for this example correspond to Figure 4. Assume that there are three states of mastery, and unequal prior probabilities for these three states. The educational decisionmaker must provide estimates for the prior probabilities of master, $p(Mi)$. For this example let us assume the values to be $p(M1) = .5$; $p(M2) = .3$; and $p(M3) = .2$. The decisionmaker must also provide estimates for the conditional probability of getting any given test item right, given each mastery state. Use the following values as the conditional probability of getting an item right, given a mastery state: $p(1|M1) = .8$; $p(1|M2) = .6$; $p(1|M3) = .5$. The conditional probabilities of getting an item wrong given a mastery state are $p(0|M1) = .2$; $p(0|M2) = .4$; and $p(0|M3) = .5$.

First we need to calculate the probability that an item is answered correctly. For the overall population,

$$p(tj = \text{correct}) = \sum_{i=1}^{S} p(Mi)p(tj = \text{correct}|Mi) = (.5)(.8)$$

$$+ (.3)(.6) + (.2)(.5) = .68.$$

Likewise,

$$p(tj = \text{wrong}) = \sum_{i=1}^{S} p(Mi)p(tj = \text{wrong}|Mi)$$

$$= (.5)(.2) + (.3)(.4) + (.2)(.5) = .32.$$

We also need to obtain the set of conditional probabilities for the different mastery states, given that an individual item was responded to either correctly or wrongly. The general equation is

$$p(Mi|tj) = \frac{p(Mi)p(tj|Mi)}{p(tj)} .$$

Substituting the above values yields

$$p(M1|tj = \text{correct}) = (.5)(.8) \div .68 = .588;$$
$$p(M2|tj = \text{correct}) = (.3)(.6) \div .68 = .265; \text{ and}$$
$$p(M3|tj = \text{correct}) = (.2)(.5) \div .68 = .147.$$

(Note that the sum equals 1.0.) Finally,

$$p(M1|tj = \text{wrong}) = (.5)(.2) \div .32 = .3125;$$
$$p(M2|tj = \text{wrong}) = (.3)(.4) \div .32 = .375; \text{ and}$$
$$p(M3|tj = \text{wrong}) = (.2)(.5) \div .32 = .3125.$$

If 6 items were answered correctly on a 10-item criterion-referenced test, the following $\prod_{j=1}^{N} p(Mi|tj)$ values result:

$$M1 = 3.9 \times 10^{-4};\ M2 = 6.8 \times 10^{-6};\ M3 = 9.6 \times 10^{-8}.$$

Finally, the general Bayesian formula yields the conditional probability for <u>each</u> mastery state given the total test score. For example,

$$p(Mi|T) = \frac{(3.9 \times 10^{-4})}{(.5)^9\left[\frac{(3.9 \times 10^{-4})}{(.5)^9} + \frac{(6.8 \times 10^{-6})}{(.3)^9} + \frac{(9.6 \times 10^{-8})}{(.2)^9}\right]} = .272.$$

Similar calculations yield $p(M2|T) = .473$ and $p(M3|T) = .254$.

In order to combine mastery states M2 and M3 into a <u>single</u> mastery state (which could represent combining the two degrees of nonmastery, Figure 4, Graph D), the following calculations are required. The values for $p(M1)$ and $\prod_{j=1}^{N} p(M1|tj)$ remain the same, .5 and 3.9×10^{-4}, respectively. The new nonmastery state (M2') occurs as a result of combining the previous states M2 and M3. Hence,

$$p(M2') = p(M2) + p(M3) = .3 + .2 = .5,$$

$$p(M2'|tj = \text{correct}) = p(M2|tj = \text{correct}) + p(M3|tj = \text{correct})$$
$$= .265 + .147 = .412, \text{ and}$$

$$p(M2'|tj = \text{wrong}) = p(M2|tj = \text{wrong}) + p(M3|tj = \text{wrong})$$
$$= .375 + .3125 = .6875.$$

Calculation of $\prod_{j=1}^{N} p(M2'|tj)$ yields

$$1.09 \times 10^{-3}.$$

Entering these new values into the general Bayesian Formula, the following values of $p(M1'|T)$ and $p(M2'|T)$ are obtained:

$$p(M1'|T) = \frac{3.9 \times 10^{-4}}{(.5)^9 \left[\frac{(3.9 \times 10^{-4})}{(.5)^9} + \frac{(1.09 \times 10^{-3})}{(.5)^9} \right]} = .264,$$

$$p(M2'|T) = \frac{1.09 \times 10^{-3}}{(.5)^9 \left[\frac{(3.9 \times 10^{-4})}{(.5)^9} + \frac{(1.09 \times 10^{-3})}{(.5)^9} \right]} = .736.$$

Some interesting properties of the model emerge when an alternative procedure for combining mastery groups is used. Note that to combine two mastery states it is <u>not</u> necessary to calculate new values for $p(1|M2')$ and $p(0|M2')$. However, it is possible to show that these values are weighted averages of $p(1|M2)$ and $p(1|M3)$, and $p(0|M2)$ and $p(0|M3)$, respectively, where the weights are the relative proportions of the new state accounted for by each of the previous states. The calculations follow.

Since $p(M2) = .3$ and $p(M3) = .2$, state M2 accounts for 60% and M3 accounts for 40% of the new state M2'. Hence, the value of

$$p(1|M2') = (.6)p(1|M2) + (.4)p(1|M3) = (.6)(.6) + (.4)(.5) = .56 \text{ and}$$

$$p(0|M2') = (.6)p(0|M2) + (.4)p(0|M3) = (.6)(.4) + (.4)(.5) = .44.$$

Using these new values,

$$p(tj = \text{correct}) = p(M1')p(1|M1') + p(M2')p(1|M2')$$
$$= (.5)(.8) + (.5)(.56) = .68 \text{ and}$$

$$p(tj = \text{wrong}) = p(M1')p(0|M1') + p(M2')p(0|M2')$$
$$= (.5)(.2) + (.5)(.44) = .32.$$

Finally, $p(M2'|1)$ and $p(M2'|0)$ may be calculated.

$$p(M2'|1) = \frac{p(M2')p(1|M2')}{p(1)} = \frac{(.5)(.56)}{.68} = .412,$$

and

$$p(M2'|0) = \frac{p(M2')p(0|M2')}{p(0)} = \frac{(.5)(.44)}{.32} = .6875.$$

These values are the same as those obtained by the simple addition procedure shown above.

This exercise serves to illustrate the effect of combining two mastery states. Combining states M2 and M3 creates, in effect, a new description of the examinee population in which only two mastery states are hypothesized. The parameter estimates for the new states in this example, are

$$p(M1) = .5 \qquad p(M2) = .5$$
$$p(1|M1) = .8 \qquad p(1|M2) = .56.$$

In choosing to combine groups, the decisionmaker must consider whether a two-state description of the population with parameter estimates such as those above is a better representation than the original three-state descriptions with parameter estimates.

$$p(M1) = .5,\ p(M2) = .3,\ p(M3) = .2,$$
$$p(1|M1) = .8,\ p(1|M2) = .6,\ p(1|M3) = .5.$$

ARI Distribution List

4 OASD (M&RA)
2 HQDA (DAMI-CSZ)
1 HQDA (DAPE-PBR
1 HQDA (DAMA-AR)
1 HQDA (DAPE-HRE-PO)
1 HQDA (SGRD-ID)
1 HQDA (DAMI-DOT-C)
1 HQDA (DAPC-PMZ-A)
1 HQDA (DACH-PPZ-A)
1 HQDA (DAPE-HRE)
1 HQDA (DAPE-MPO-C)
1 HQDA (DAPE-DW)
1 HQDA (DAPE-HRL)
1 HQDA (DAPE-CPS)
1 HQDA (DAFD-MFA)
1 HQDA (DARD-ARS-P)
1 HQDA (DAPC-PAS-A)
1 HQDA (DUSA-OR)
1 HQDA (DAMO-RQR)
1 HQDA (DASG)
1 HQDA (DA10-PI)
1 Chief, Consult Div (DA-OTSG), Adelphi, MD
1 Mil Asst. Hum Res, ODDR&E, OAD (E&LS)
1 HQ USARAL, APO Seattle, ATTN: ANAGP-R
1 HQ First Army, ATTN: AFKA-OI-TI
2 HQ Fifth Army, Ft Sam Houston
1 Dir, Army Stf Studies Ofc, ATTN: OAVCSA (DSP)
1 Ofc Chief of Stf, Studies Ofc
1 DCSPER, ATTN: CPS/OCP
1 The Army Lib, Pentagon, ATTN: RSB Chief
1 The Army Lib, Pentagon, ATTN: ANRAL
1 Ofc, Asst Sect of the Army (R&D)
1 Tech Support Ofc, OJCS
1 USASA, Arlington, ATTN: IARD-T
1 USA Rsch Ofc, Durham, ATTN: Life Sciences Dir
2 USARIEM, Natick, ATTN: SGRD-UE-CA
1 USATTC, Ft Clayton, ATTN: STETC-MO-A
1 USAIMA, Ft Bragg, ATTN: ATSU-CTD-OM
1 USAIMA, Ft Bragg, ATTN: Marquat Lib
1 US WAC Ctr & Sch, Ft McClellan, ATTN: Lib
1 US WAC Ctr & Sch, Ft McClellan, ATTN: Trng Dir
1 USA Quartermaster Sch, Ft Lee, ATTN: ATSM-TE
1 Intelligence Material Dev Ofc, EWL, Ft Holabird
1 USA SE Signal Sch, Ft Gordon, ATTN: ATSO-EA
1 USA Chaplain Ctr & Sch, Ft Hamilton, ATTN: ATSC-TE-RD
1 USATSCH, Ft Eustis, ATTN: Educ Advisor
1 USA War College, Carlisle Barracks, ATTN: Lib
2 WRAIR, Neuropsychiatry Div
1 DLI, SDA, Monterey
1 USA Concept Anal Agcy, Bethesda, ATTN: MOCA-WGC
1 USA Concept Anal Agcy, Bethesda, ATTN: MOCA-MR
1 USA Concept Anal Agcy, Bethesda, ATTN: MOCA-JF
1 USA Artic Test Ctr, APO Seattle, ATTN: STEAC-MO-ASL
1 USA Artic Test Ctr, APO Seattle, ATTN: AMSTE-PL-TS
1 USA Armament Cmd, Redstone Arsenal, ATTN: ATSK-TEM
1 USA Armament Cmd, Rock Island, ATTN: AMSAR-TDC
1 FAA-NAFEC, Atlantic City, ATTN: Library
1 FAA-NAFEC, Atlantic City, ATTN: Hum Engr Br
1 FAA Aeronautical Ctr, Oklahoma City, ATTN: AAC-44D
2 USA Fld Arty Sch, Ft Sill, ATTN: Library
1 USA Armor Sch, Ft Knox, ATTN: Library
1 USA Armor Sch, Ft Knox, ATTN: ATSB-DI-E
1 USA Armor Sch, Ft Knox, ATTN: ATSB-DT-TP
1 USA Armor Sch, Ft Knox, ATTN: ATSB-CD-AD
2 HQUSACDEC, Ft Ord, ATTN: Library
1 HQUSACDEC, Ft Ord, ATTN: ATEC-EX-E-Hum Factors
2 USAEEC, Ft Benjamin Harrison, ATTN: Library
1 USAPACDC, Ft Benjamin Harrison, ATTN: ATCP-HR
1 USA Comm-Elect Sch, Ft Monmouth, ATTN: ATSN-EA
1 USAEC, Ft Monmouth, ATTN: AMSEL-CT-HDP
1 USAEC, Ft Monmouth, ATTN: AMSEL-PA-P
1 USAEC, Ft Monmouth, ATTN: AMSEL-SI-CB
1 USAEC, Ft Monmouth, ATTN: C, Facl Dev Br
1 USA Materials Sys Anal Agcy, Aberdeen, ATTN: AMXSY-P
1 Edgewood Arsenal, Aberdeen, ATTN: SAREA-BL-H
1 USA Ord Ctr & Sch, Aberdeen, ATTN: ATSL-TEM-C
2 USA Hum Engr Lab, Aberdeen, ATTN: Library/Dir
1 USA Combat Arms Trng Bd, Ft Benning, ATTN: Ad Supervisor
1 USA Infantry Hum Rsch Unit, Ft Benning, ATTN: Chief
1 USA Infantry Bd, Ft Benning, ATTN: STEBC-TE-T
1 USASMA, Ft Bliss, ATTN: ATSS-LRC
1 USA Air Def Sch, Ft Bliss, ATTN: ATSA-CTD-ME
1 USA Air Def Sch, Ft Bliss, ATTN: Tech Lib
1 USA Air Def Bd, Ft Bliss, ATTN: FILES
1 USA Air Def Bd, Ft Bliss, ATTN: STEBD-PO
1 USA Cmd & General Stf College, Ft Leavenworth, ATTN: Lib
1 USA Cmd & General Stf College, Ft Leavenworth, ATTN: ATSW-SE-L
1 USA Cmd & General Stf College, Ft Leavenworth, ATTN: Ed Advisor
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: DepCdr
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: CCS
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: ATCASA
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: ATCACO-E
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: ATCACC-C
1 USAECOM, Night Vision Lab, Ft Belvoir, ATTN: AMSEL-NV-SD
1 USA Computer Sys Cmd, Ft Belvoir, ATTN: Tech Library
1 USAMERDC, Ft Belvoir, ATTN: STSFB-DQ
1 USA Eng Sch, Ft Belvoir, ATTN: Library
1 USA Topographic Lab, Ft Belvoir, ATTN: ETL-TD-S
1 USA Topographic Lab, Ft Belvoir, ATTN: STINFO Center
1 USA Topographic Lab, Ft Belvoir, ATTN: ETL-GSL
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: CTD-MS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATS-CTD-MS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-TE
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-TEX-GS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-CTS-OR
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-CTD-DT
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-CTD-CS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: DAS/SRD
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-TEM
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: Library
1 CDR, HQ Ft Huachuca, ATTN: Tech Ref Div
2 CDR, USA Electronic Prvg Grd, ATTN: STEEP-[illegible]-S
1 CDR, Project MASSTER, ATTN: Tech Info Center
1 Hq MASSTER, USATRADOC, LNO
1 Research Institute, HQ MASSTER, Ft Hood
1 USA Recruiting Cmd, Ft Sheridan, ATTN: USARCPM-P
1 Senior Army Adv., USAFAGOD/TAC, Elgin AF Aux Fld No. 9
1 HQ USARPAC, DCSPER, APO SF 96558, ATTN: GPPE-SE
1 Stimson Lib, Academy of Health Sciences, Ft Sam Houston
1 Marine Corps Inst., ATTN: Dean-MCI
1 HQUSMC, Commandant, ATTN: Code MTMT 51
1 HQUSMC, Commandant, ATTN: Code MPI-20
2 USCG Academy, New London, ATTN: Admission
2 USCG Academy, New London, ATTN: Library
1 USCG Training Ctr, NY, ATTN: CO
1 USCG Training Ctr, NY, ATTN: Educ Svc Ofc
1 USCG, Psychol Res Br, DC, ATTN: GP 1/62
1 HQ Mid-Range Br, MC Det, Quantico, ATTN: P&S Div

1 US Marine Corps Liaison Ofc, AMC, Alexandria, ATTN: AMCGS-F
1 USATRADOC, Ft Monroe, ATTN: ATRO-ED
6 USATRADOC, Ft Monroe, ATTN: ATPR-AD
1 USATRADOC, Ft Monroe, ATTN: ATTS-EA
1 USA Forces Cmd, Ft McPherson, ATTN: Library
2 USA Aviation Test Bd, Ft Rucker, ATTN: STEBG-PO
1 USA Agcy for Aviation Safety, Ft Rucker, ATTN: Library
1 USA Agcy for Aviation Safety, Ft Rucker, ATTN: Educ Advisor
1 USA Aviation Sch, Ft Rucker, ATTN: PO Drawer O
1 HQUSA Aviation Sys Cmd, St Louis, ATTN: AMSAV-ZDR
2 USA Aviation Sys Test Act., Edwards AFB, ATTN: SAVTE-T
1 USA Air Def Sch, Ft Bliss, ATTN: ATSA TEM
1 USA Air Mobility Rsch & Dev Lab, Moffett Fld, ATTN: SAVDL-AS
1 USA Aviation Sch, Res Trng Mgt, Ft Rucker, ATTN: ATST-T-RTM
1 USA Aviation Sch, CO, Ft Rucker, ATTN: ATST-D-A
1 HQ, DARCOM, Alexandria, ATTN: AMXCD-TL
1 HQ, DARCOM, Alexandria, ATTN: CDR
1 US Military Academy, West Point, ATTN: Serials Unit
1 US Military Academy, West Point, ATTN: Ofc of Milt Ldrshp
1 US Military Academy, West Point, ATTN: MAOR
1 USA Standardization Gp, UK, FPO NY, ATTN: MASE-GC
1 Ofc of Naval Rsch, Arlington, ATTN: Code 452
3 Ofc of Naval Rsch, Arlington, ATTN: Code 458
1 Ofc of Naval Rsch, Arlington, ATTN: Code 450
1 Ofc of Naval Rsch, Arlington, ATTN: Code 441
1 Naval Aerospc Med Res Lab, Pensacola, ATTN: Acous Sch Div
1 Naval Aerospc Med Res Lab, Pensacola, ATTN: Code L51
1 Naval Aerospc Med Res Lab, Pensacola, ATTN: Code L5
1 Chief of Naval Ops, ATTN: Pers-OR
1 NAVAIRSTA, Norfolk, ATTN: Safety Ctr
1 Nav Oceanographic, DC, ATTN: Code 6251, Charts & Tech
1 Center of Naval Anal, ATTN: Doc Ctr
1 NavAirSysCom, ATTN: AIR-5313C
1 Nav BuMed, ATTN: 713
1 NavHelicopterSubSqua 2, FPO SF 96601
1 AFHRL (FT) William AFB
1 AFHRL (TT) Lowry AFB
1 AFHRL (AS) WPAFB, OH
2 AFHRL (DOJZ) Brooks AFB
1 AFHRL (DOJN) Lackland AFB
1 HQUSAF (INYSD)
1 HQUSAF (DPXXA)
1 AFVTG (RD) Randolph AFB
3 AMRL (HE) WPAFB, OH
2 AF Inst of Tech, WPAFB, OH, ATTN: ENE/SL
1 ATC (XPTD) Randolph AFB
1 USAF AeroMed Lib, Brooks AFB (SUL-4), ATTN: DOC SEC
1 AFOSR (NL), Arlington
1 AF Log Cmd, McClellan AFB, ATTN: ALC/DPCRB
1 Air Force Academy, CO, ATTN: Dept of Bel Sen
5 NavPers & Dev Ctr, San Diego
2 Navy Med Neuropsychiatric Rsch Unit, San Diego
1 Nav Electronic Lab, San Diego, ATTN: Res Lab
1 Nav TrngCen, San Diego, ATTN: Code 9000-Lib
1 NavPostGraSch, Monterey, ATTN: Code 55As
1 NavPostGraSch, Monterey, ATTN: Code 2124
1 NavTrngEquipCtr, Orlando, ATTN: Tech Lib
1 US Dept of Labor, DC, ATTN: Manpower Admin
1 US Dept of Justice, DC, ATTN: Drug Enforce Admin
1 Nat Bur of Standards, DC, ATTN: Computer Info Section
1 Nat Clearing House for MH-Info, Rockville
1 Denver Federal Ctr, Lakewood, ATTN: BLM
12 Defense Documentation Center
4 Dir Psych, Army Hq, Russell Ofcs, Canberra
1 Scientific Advsr, Mil Bd, Army Hq, Russell Ofcs, Canberra
1 Mil and Air Attache, Austrian Embassy
1 Centre de Recherche Des Facteurs, Humaine de la Defense Nationale, Brussels
2 Canadian Joint Staff Washington
1 C/Air Staff, Royal Canadian AF, ATTN: Pers Std Anal Br
3 Chief, Canadian Def Rsch Staff, ATTN: C/CRDS(W)
4 British Def Staff, British Embassy, Washington
1 Def & Civil Inst of Enviro Medicine, Canada
1 AIR CRESS, Kensington, ATTN: Info Sys Br
1 Militaerpsykologisk Tjeneste, Copenhagen
1 Military Attache, French Embassy, ATTN: Doc Sec
1 Medecin Chef, C.E.R.P.A.-Arsenal, Toulon/Naval France
1 Prin Scientific Off, Appl Hum Engr Rsch Dir, Ministry of Defense, New Delhi
1 Pers Rsch Ofc Library, AKA, Israel Defense Forces
1 Ministerie van Defensie, DOOP/KL Afd Sociaal Psychologische Zaken, The Hague, Netherlands

52